UNITED STATES PATENT APPLICATION

FOR

PRIORITY AWARE MAC FLOW CONTROL

INVENTORS:

Nitin Jain, a citizen of Canada
Rajkumar Jalan, a citizen of the United States

ASSIGNED TO:

Foundry Networks, Inc., a Delaware Corporation

PREPARED BY:

THELEN, REID & PRIEST LLP
P.O. BOX 640640
SAN JOSE, CA 95164-0640
TELEPHONE: (408) 292-5800
FAX: (408) 287-8040

Attorney Docket Number: 034103-0026

Client Docket Number: FOUND-0026

SPECIFICATION

TITLE OF INVENTION

PRIORITY AWARE MAC FLOW CONTROL

COPYRIGHT NOTICE

[0001] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyright rights whatsoever.

FIELD OF THE INVENTION

[0002] The present invention relates to the field of switching in a computer network. More particularly, the present invention relates to the controlling the flow of packets through Media Access Control (MAC) layer switching, while honoring the priority levels of the packets.

BACKGROUND OF THE INVENTION

[0003] A switch is a device that provides a switching function (i.e., determines a physical path) in a data communications network. Switching may often involve transferring information, such as digital data packets or frames, among entities of the network. Switching is accomplished by examining data on one or more of the network layers. One such type of switching is media access control (MAC) address-based switching, which involves switching in the data link layer. The data link layer is also commonly referred to as layer 2 of the OSI Reference Model.

Specifically, switching may occur through Ethernet and/or Gigabit Ethernet on full duplex ports for layer 2 networks.

[0004] During switching, there is often a need for flow control of packets, in case of network outages or if a network device does not have enough resources to handle the received traffic. In a full duplex network, a receiver can signal to the transmitter to stop sending the traffic if it does not have enough resources to handle the traffic. The IEEE 802.3x Specification (now part of 802.3, Annex 31B), uses PAUSE frames for a device to signal another device. The specialized MAC control PAUSE frames according to IEEE 802.3x are depicted in FIG. 1. Each frame 100 comprises a destination address 102 (6 bytes), a source address 104 (6 bytes), a type/length field 106 (2 bytes), an opcode 108 (2 bytes), a pause time field (2 bytes) 110 and 42 bytes of reserved empty space 112. When a frame is transmitted, it is preceded by a 7 byte preamble and 1 byte Start-Frame-Delimiter, and then followed by a 4 byte frame check sequence.

[0005] The PAUSE frame causes any device receiving it to stop forwarding traffic to the requested device for the specified period of time. The hope is that when that period of time is up, the device has resources available for the traffic.

[0006] This flow control mechanism, however, does not discriminate among the incoming frames - it applies to all incoming frames to a device. In certain systems, however, frames may be prioritized. For example, voice data may have a high priority level as it is extremely time sensitive, whereas text data may have a low priority level. Furthermore, certain subscriber's traffic may be afforded higher priority than others. The prior art flow control mechanism, however, violates these priorities by simply ceasing all incoming transmissions. This can even defeat the purpose of flow control in the first place, by deteriorating network throughput, causing more transmissions, and a compounding of the problem.

[0007] What is needed is a mechanism wherein the MAC can take the action of the flow control and apply it in a way that takes into account the priority of the frames.

[0008] Furthermore, currently PAUSE frames are sent out as untagged and only have significance on a single link. FIG. 2 is a diagram illustrating a typical system utilizing PAUSE frames. Here, the device that transmits the PAUSE frame 200 wishes to cause another device 202 to hold off on transmitting frames for a time. The other device 202 processes the PAUSE frame it receives but does nothing further with the frame itself.

[0009] However, in the metro Ethernet environment, clients and servers may not be directly connected, but rather connected over several hops. FIG. 3 is a diagram illustrating a typical metro Ethernet environment. The transmitting device 300 in a first VLAN is separated from the receiving device 302, also in the first VLAN, by several hops 304-314, which are typically switches or hubs. Currently, the point-to-point nature of the PAUSE mechanism prevents the receiving device 302 from receiving the PAUSE frame, because the first hop 304 processes the frame without forwarding it. What is needed is a mechanism to extend the PAUSE frame solution to Virtual Local Area Networks (VLANs) across multiple hops. What is also needed is a mechanism that would allow the traffic flow in a specific VLAN to be paused, without pausing traffic flow in other VLANs.

BRIEF DESCRIPTION

[0010] Solutions are provided that allow a network device to apply flow control on the MAC layer while taking into account the priority of the frame of traffic. This may be accomplished by generating a frame indicating that traffic flow should be paused, while utilizing a new opcode value, or alternatively by utilizing a new type/length value (possibly combined with a new opcode value). A receiving device may then examine the fields of the frame to determine whether it should it should use priority-based pausing, and then examine other fields to determine which priority-levels to pause and for how long. This allows for improved efficiency in flow control on the MAC layer.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying drawings, which are incorporated into and constitute a part of this specification, illustrate one or more embodiments of the present invention and, together with the detailed description, serve to explain the principles and implementations of the invention.

[0012] In the drawings:

FIG. 1 is a diagram illustrating MAC control PAUSE frames according to IEEE 802.3x.

FIG. 2 is a diagram illustrating a typical system utilizing PAUSE frames.

FIG. 3 is a diagram illustrating a typical metro Ethernet environment.

FIG. 4 is a diagram illustrating a PAUSE frame format in accordance with an embodiment of the present invention.

FIG. 5 is a diagram illustrating an example PAUSE frame in accordance with an embodiment of the present invention.

FIG. 6 is a flow diagram illustrating a method for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with an embodiment of the present invention.

FIG. 7 is a flow diagram illustrating a method for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with another embodiment of the present invention.

FIG. 8 is a flow diagram illustrating a method for handling a frame in a network with traffic flow having varying priority levels, in accordance with an embodiment of the present invention.

FIG. 9 is a flow diagram illustrating a method for handling a frame in a network with traffic flow having varying priority levels, in accordance with another embodiment of the present invention.

FIG. 10 is a block diagram illustrating an apparatus for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with an embodiment of the present invention.

FIG. 11 is a block diagram illustrating an apparatus for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with another embodiment of the present invention.

FIG. 12 is a block diagram illustrating an apparatus for handling a frame in a network with traffic flow having varying priority levels, in accordance with an embodiment of the present invention.

FIG. 13 is a block diagram illustrating an apparatus for handling a frame in a network with traffic flow having varying priority levels, in accordance with another embodiment of the present invention.

DETAILED DESCRIPTION

[0013] Embodiments of the present invention are described herein in the context of a system of computers, servers, and software. Those of ordinary skill in the art will realize that the following detailed description of the present invention is illustrative only and is not intended to be in any way limiting. Other embodiments of the present invention will readily suggest themselves to such skilled persons having the benefit of this disclosure. Reference will now be made in detail to implementations of the present invention as illustrated in the accompanying drawings. The same reference indicators will be used throughout the drawings and the following detailed description to refer to the same or like parts.

[0014] In the interest of clarity, not all of the routine features of the implementations described herein are shown and described. It will, of course, be appreciated that in the development of any such actual implementation, numerous implementation-specific decisions must be made in order to achieve the developer's specific goals, such as compliance with application- and business-related constraints, and that these specific goals will vary from one implementation to another and from one developer to another. Moreover, it will be appreciated that such a development effort might be complex and time-consuming, but would nevertheless be a routine undertaking of engineering for those of ordinary skill in the art having the benefit of this disclosure.

[0015] In accordance with the present invention, the components, process steps, and/or data structures may be implemented using various types of operating systems, computing platforms, computer programs, and/or general purpose machines. In addition, those of ordinary skill in the art will recognize that devices of a less general purpose nature, such as hardwired devices, field programmable gate arrays (FPGAs), application specific integrated circuits (ASICs), or the like, may also be used without departing from the scope and spirit of the inventive concepts disclosed herein. Furthermore, the present invention is described in the context of a switch. However, one of ordinary skill in the art will recognize that the term switch should be read broadly, so as to include any device that directs packets, including a router and a gateway.

[0016] The present invention provides mechanisms to allow a network device to apply flow control on the MAC layer while taking into account the priority of the frames of traffic.

[0017] Two mechanisms are described herein for applying flow control on a MAC layer for packets having a priority value. One of ordinary skill in the art will recognize that the specifics of these mechanisms are merely examples and should not be read as limiting. In one embodiment of the present invention, a different opcode value along with a new field may be utilized in the PAUSE frame in order to indicate how to handle frames of varying priorities. The new field may be termed a priority mask, and may be used to identify to which priority to apply the PAUSE command. Thus, FIG. 4 is a diagram illustrating a PAUSE frame format in accordance with an embodiment of the present invention. Destination address 400, source address 402 and type/length 404 may remain the same as the typical PAUSE frame. The opcode field 406 may contain a different value.

[0018] The new priority mask field 408 may be two bytes long, however the first byte may be unused in systems having 8 or fewer possible priority levels. In this embodiment, each bit of the second byte may correspond to a different priority level. Therefore, bit 0 might correspond to a priority level of 0, bit 1 to a priority level of 1, etc. The presence of any bit signals the traffic with the specific priority to be paused.

[0019] In an embodiment of the present invention, the pause time field 410 may be extended to 16 bytes, to allow for each priority level to have a different pause time. This may be utilized only when it is desired to have varying pause times - if it is more desirable in a specific instance to have a single pause time for all paused traffic, the only pause time field may be used. The new pause time field allows for 8 2-byte values for pause time. For example, as depicted in FIG. 5, if it is desired for all traffic with priorities of 0, 1, and 2 to be paused, with the pause time value of traffic with priority 0 being 7, traffic with priority 1 being 5, and traffic with priority 2 being 3, then the priority mask 500 may be set at 00000111, and the pause time array 502 set at zero for each of the first 5 2-byte entries, the sixth entry being set at 0x3, the seventh at 0x5, and the eighth at 0x7.

[0020] Typically, the PAUSE frame utilizes an opcode value of 1. In an embodiment of the present invention, an opcode value of 2 may indicate the presence of the priority mask field - thus the receiving device would pause traffic with a priority value indicated by the priority mask. The pausing in this instance would be for a set time for all priorities, thus using only a single value in the pause time field.

[0021] An opcode value of 3, then, may indicate the presence of both the priority mask field and the new pause time field, thus indicating to the receiving device that it should pause traffic with a priority value indicated by the priority mask, for time periods as specified in the new pause time field.

[0022] In another embodiment of the present invention, a new type/length value may be used. This embodiment is beneficial when encountering devices utilizing older MAC standards, which may not be able to understand the new opcode values described above. Typically, the value "8808" is utilized in the type/length field to indicate a PAUSE frame. In this embodiment, the value "8809" may be used, for example, to indicate that this is a PAUSE frame that handles priority. The opcode field may then be used to indicate whether or not all the traffic priority levels utilize the same pause time - rather than values of 2 and 3 they may be, for example, 1 and 2. Otherwise, the frame format described in FIG. 5 may be utilized in this embodiment as is. Thus, the presence of "8809" in the type/length field along with a value of 1 in the opcode field would indicate the presence of the priority mask field and that the receiving device should pause traffic with the corresponding priority value(s) for a set, single period of time set in the pause time field. The presence of "8809" in the type/length field along with a value of 2 in the opcode field would indicate the presence of the priority mask field and the new pause time field, thus indicating to the receiving device that it should pause traffic with a priority value indicated by the priority mask, for time periods as specified in the new pause time field.

[0023] FIG. 6 is a flow diagram illustrating a method for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with an embodiment of the present invention. At 600, a value signifying that the frame indicates that traffic flow should be paused may be placed in a type/length field in the frame. This may be a value identical to that of standard PAUSE frames, for example. At 602, a value signifying that traffic flow should be paused or not paused according to its priority level may be placed in an opcode field in the frame. This value may also indicate whether the pausing will be for time indicated by a pause time field in the frame without regard for the priority level (if the same pause time for each priority level is desired), or whether the pausing will be for times corresponding to each priority level indicated by the pause time field (if independent pause times for each priority level are desired). If the latter, then at 604, a separate value for each possible priority level may be placed in the pause time field, the separate value indicating an independent pause time for each corresponding priority level. The pause time field in that case may be equal in size to the pause time field in a standard PAUSE frame multiplied by the number of possible priority levels. These opcode values may be values not used by standard PAUSE frames in the opcode field. At 606, a priority mask field may be created in the frame. At 608, a value signifying which priority levels should be paused may be placed in the priority mask field in the frame.

[0024] FIG. 7 is a flow diagram illustrating a method for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with another embodiment of the present invention. At 700, a value signifying that the frame indicates that traffic flow should be paused or not paused according to its priority level may be placed in a type/length field in the frame. This may be a value unused in standard PAUSE frames, for example. At 702, a value signifying that the pausing will be for a time indicated by a pause time field in the frame without regard for the priority level (if the same pause time for each priority level is desired), or that the pausing will be for times corresponding to each priority level indicated by the pause time field (if independent pause times for each priority level are desired) may be placed in an opcode field in the frame. If the latter, then at 704, a separate value for each possible priority level may be placed in the pause time field, the separate value indicating an independent pause time for each corresponding priority level. The pause time field in that case may be equal in size to the pause time field in a standard PAUSE frame multiplied by the number of possible priority levels. At 706, a priority mask field may be created in the frame. At 708, a value signifying which priority levels should be paused may be placed in the priority mask field in the frame.

[0025] FIG. 8 is a flow diagram illustrating a method for handling a frame in a network with traffic flow having varying priority levels, in accordance with an embodiment of the present invention. At 800, a value in a type/length field in the frame may be examined to determine if it signifies that the frame indicates that traffic flow should be paused to a network device. If it does, then at 802, a value in an opcode field in the frame may be examined to determine if it signifies that traffic flow should be paused or not paused according to its priority level. If this is also true, then at 804, traffic flow with priority levels corresponding to levels signified by a value in a priority mask field in the frame may be paused. At 802, the examining of the opcode field may include examining it to determine if it also signifies that the pausing will be for a time indicated by a pause time field in the frame without regard to priority level or whether the pausing will be for times corresponding to each priority level indicated by the pause time field. If the former, then at 804, the pausing may include pausing the traffic flow for a time period indicated by the pause time field in the frame without regard to priority level. If the latter, then at 804, the pausing may include pausing the traffic flow for time periods indicated by times corresponding to each priority level in the pause time field. These times may be a separate value for each possible priority level indicating an independent pause time for each corresponding priority level.

[0026] FIG. 9 is a flow diagram illustrating a method for handling a frame in a network with traffic flow having varying priority levels, in accordance with another embodiment of the present invention. At 900, a value in a type/length field in the frame may be examined to determine if it signifies that the frame indicates that traffic flow should be paused to a network device and, at 902, if it signifies that traffic flow should be paused or not paused according to its priority level. If both are true, then at 904, a value in an opcode field in the frame may be examined to determine if it signifies that the pausing will be for a time indicated by a pause time field in the frame without regard to priority level or whether the pausing will be for times corresponding to each priority level indicated by the pause time field. If the former, then at 906, the traffic flow with priority levels corresponding to levels signified by a value in a priority mask field in the frame may be paused for a time period indicated by the pause time field in the frame without regard to priority level. If the latter, then at 908, the traffic flow with priority levels corresponding to levels signified by a value in a priority mask field in the frame may be paused for time periods indicated by times corresponding to each priority level in the pause time field. These times may be a separate value for each possible priority level indicating an independent pause time for each corresponding priority level.

[0027] FIG. 10 is a block diagram illustrating an apparatus for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with an embodiment of the present invention. A pause traffic flow value-to-type/length field placer 1000 may place a value signifying that the frame indicates that traffic flow should be paused in a type/length field in the frame. This may be a value identical to that of standard PAUSE frames, for example. A priority level based pause traffic flow value-to-opcode field placer 1002 coupled to the pause traffic flow value-to-type/length field placer 1000 may place a value signifying that traffic flow should be paused or not paused according to its priority level in an opcode field in the frame. This value may also indicate whether the pausing will be for time indicated by a pause time field in the frame without regard for the priority level (if the same pause time for each priority level is desired), by using a pause time without regard for priority level value-to-opcode field placer 1004, or whether the pausing will be for times corresponding to each priority level indicated by the pause time field (if independent pause times for each priority level are desired), by using a pause times corresponding to priority level value-to-opcode field placer 1006. If the latter, then a priority level separate value-to-pause time field placer 1008 coupled to the priority level based pause traffic flow value-to-opcode field placer 1002 may place a separate value for each possible priority level in the pause time field, the separate value indicating an independent pause time for each corresponding priority level. The pause time field in that case may be equal in size to the pause time field in a standard PAUSE frame multiplied by the number of possible priority levels. These opcode values may be values not used by standard PAUSE frames in the opcode field. A priority mask field creator 1010 coupled to the priority level based pause traffic flow value-to-opcode field placer 1002 may create a priority mask field in the frame. A paused priority level value-to-priority mask field placer 1012 coupled to the priority mask field creator 1010 may place a value signifying which priority levels should be paused in the priority mask field in the frame.

[0028] FIG. 11 is a block diagram illustrating an apparatus for generating a frame indicating that traffic flow should be paused to a network device, the traffic flow having varying priority levels, in accordance with another embodiment of the present invention. A priority level based pause traffic flow value-to-type/length field placer 1100 may place a value signifying that the frame indicates that traffic flow should be paused or not paused according to its priority level in a type/length field in the frame. This may be a value unused in standard PAUSE frames, for example. A pause time without regard for priority level value-to-opcode field placer 1102 coupled to the priority level based pause traffic flow value-to-type/length field placer 1100 may place a value in the opcode field signifying that the pausing will be for a time indicated by a pause time field in the frame without regard for the priority level if the same pause time for each priority level is desired. Alternatively, a pause times corresponding to priority level value-to-opcode field placer 1104 coupled to the priority level based pause traffic flow value-to-type/length field placer 1100 may place a value in the opcode field signifying that the pausing will be for times corresponding to each priority level indicated by the pause time field if independent pause times for each priority level are desired. If the latter, then a priority level separate value-to-pause time field placer 1106 coupled to the pause times corresponding to priority level value-to-opcode field placer 1104 may place a separate value for each possible priority level in the pause time field, the separate value indicating an independent pause time for each corresponding priority level. The pause time field in that case may be equal in size to the pause time field in a standard PAUSE frame multiplied by the number of possible priority levels. A priority mask field creator 1108 coupled to the priority level based pause traffic flow value-to-type/length field placer 1100 may create a priority mask field in the frame. A paused priority level value-to-priority mask field placer 1110 coupled to the priority mask field creator 1108 may place a value signifying which priority levels should be paused in the priority mask field in the frame.

[0029] FIG. 12 is a block diagram illustrating an apparatus for handling a frame in a network with traffic flow having varying priority levels, in accordance with an embodiment of the present invention. A type/length field value examiner 1200 may examine a value in a type/length field in the frame to determine if it signifies that the frame indicates that traffic flow should be paused to a network device. If it does, then an opcode field value examiner 1202 coupled to the type/length field value examiner 1200 may examine a value in an opcode field in the frame to determine if it signifies that traffic flow should be paused or not paused according to its priority level. If this is also true, then a priority level traffic flow pauser 1204 coupled to the opcode field value examiner 1202 may pause traffic flow with priority levels corresponding to levels signified by a value in a priority mask field in the frame. The examining of the opcode field may include examining it to determine if it also signifies that the pausing will be for a time indicated by a pause time field in the frame without regard to priority level or whether the pausing will be for times corresponding to each priority level indicated by the pause time field. If the former, then the pausing may include pausing the traffic flow for a time period indicated by the pause time field in the frame without regard to priority level. If the latter, then the pausing may include pausing the traffic flow for time periods indicated by times corresponding to each priority level in the pause time field. These times may be a separate value for each possible priority level indicating an independent pause time for each corresponding priority level.

[0030] FIG. 13 is a block diagram illustrating an apparatus for handling a frame in a network with traffic flow having varying priority levels, in accordance with another embodiment of the present invention. A type/length field value examiner 1300 may examine a value in a type/length field in the frame to determine if it signifies that the frame indicates that traffic flow should be paused to a network device and if it signifies that traffic flow should be paused or not paused according to its priority level. If both are true, then a priority level traffic flow pauser 1302 coupled to the type/length field value examiner 1300 may pause the traffic flow with priority levels corresponding to levels signified by a value in a priority mask field in the frame.

[0031] While embodiments and applications of this invention have been shown and described, it would be apparent to those skilled in the art having the benefit of this disclosure that many more modifications than mentioned above are possible without departing from the inventive concepts herein. The invention, therefore, is not to be restricted except in the spirit of the appended claims.